Treball de Fi de Grau

Grau en Enginyeria en Tecnologies Industrials

Machine Learning aplicat al Data Warehouse tradicional

MEMÒRIA

Autor: Adrià Recort I Fernandez
Director: Daniela Tost Padreny

Convocatòria: Gener 2020

Escola Tècnica Superior
d'Enginyeria Industrial de Barcelona

Resum

Aquest treball sorgeix com a proposta de l'alumne a l'empresa en que es van realitzar les pràctiques curriculars. Es pretén realitzar una introducció i aplicació de les tècniques de Machine learning a la base de dades d'una empresa real amb l'objectiu final d'aconseguir resultats sobre un cas d'ús real (som capaços de predir si els clients de l'empresa tornaran a comprar productes de la companyia?).

L'objectiu principal del treball es determinar fins a quin punt, una inversió mínima en un projecte desenvolupat a mode de prova de concepte permet a una empresa decidir si aquest tipus de tècniques d'anàlisi de dades suposen una inversió rentable.

La metodologia a emprar començarà per comparar els pressupostos d'un projecte com seria aquest treball amb d'altres ofertes reals del mercat, s'analitzarà l'estat de l'art en tècniques de Machine learning per a aconseguir un nivell de formació suficient com per a elaborar un prototip de predictor que satisfaci l'objectiu del projecte, s'analitzarà les dades per escollir les variables que s'utilitzaran finalment, s'implementarà el predictor i se'n analitzaran els resultats.

Per a poder realitzar l'anàlisi de comportament dels diferents models, es comparen els rendiments esperats per la definició intrínseca de cada algoritme en les diferents àrees que s'estudien de manera habitual en base al problema a resoldre tal i com s'explica al capítol 4.

El model que s'ha utilitzat finalment s'anomena XGBoost, variant del model *random forest* que iterativament augmenta en complexitat per a la presa de decisions fins a un límit marcat pel desenvolupador (en aquest cas l'estudiant) per a evitar problemes d'*overfitting* (entrenar massa el model en el conjunt d'entrenament, fent-lo molt precís en aquest però erroni en el conjunt general).

Finalment s'arriba a la conclusió que l'algoritme escollit ha arribat a proporcionar resultats amb un marge d'error acceptable i demostra el potencial que tenen les tècniques de Machine learning de cara a l'explotació de les dades a les empreses.

Sumari

SUMARI 5

1. GLOSSARI 7

2. PREFACI 9

2.1. Origen del projecte 9

2.2. Motivació 9

3. INTRODUCCIÓ 11

3.1. Context 11

3.2. Objectius del projecte 14

3.2.1. Objectiu 1 – Estat de l'art 14

3.2.2. Objectiu 2 – Anàlisi i tractament de dades 14

3.2.3. Objectiu 3 – Disseny i implementació 15

3.2.4. Objectiu 4 – Anàlisi de resultats 15

3.3. Abast del projecte 15

4. ESTAT DE L'ART 17

4.1. Tipus d'algoritmes: 17

4.1.1. Aprenentatge supervisat: 17

4.1.2. Aprenentatge sense supervisió: 18

4.1.3. Aprenentatge reforçat: 18

4.2. Ampliació dels algoritmes: 19

4.2.1. Regressió: 19

4.2.2. Arbre de decisions: 19

4.2.3. Bosc aleatori: 20

4.2.4. K-propers-veïns (KNN): 20

4.3. Elecció de l'algoritme 21

4.3.1. Parcialitat vers variància: 22

4.3.2. Complexitat de la funció i dades d'entrenament: 22

4.3.3. Dimensionalitat: 22

4.3.4. Soroll: 23

4.3.5. Altres factors rellevants: 23

4.3.6. Comparació d'algoritmes: 25

5. ANÀLISI I TRACTAMENT DE DADES 26

5.1. Estandardització 26

5.1.1. Neteja de dades 26

5.1.2. Diccionaris 27
5.2. Visualització de les dades 27
5.2.1. Enfocament al client 28
5.2.2. Enfocament al producte 31
5.2.3. Altre informació disponible 32
5.2.4. Eliminació de valors anòmals 35
5.3. Extracció del data set 35
5.3.1. Oracle SQL 35
5.3.2. SAP BO 36
6. DISSENY I IMPLEMENTACIÓ DEL PREDICTOR 39
6.1. Parametrització 39
6.2. Entrenament, validació i predicció 40
6.2.1. Primer cas pràctic 40
6.2.2. Iteració per paràmetres 41
6.2.3. Validació creuada 42
6.2.4. Predicció 43
7. ANÀLISI DE RESULTATS 44
7.1.1. Resultats de l'entrenament 44
7.1.2. Resultats de la predicció 45
7.1.3. Lliurable final 45
CONCLUSIONS 47
AGRAÏMENTS 49
BIBLIOGRAFIA 50
Referències bibliogràfiques 50
Bibliografia complementària 50

1. Glossari

Data Warehouse: Magatzem de dades. Base de dades estructurada on una empresa es dedica a recopilar tota la informació disponible de processos, serveis, clients i productes.

Documentació IT-PEP: Metodologia de documentació per a projectes informàtics que defineix uns arxius mínims i necessaris per assegurar el correcte enteniment i utilització per part de l'usuari final.

Palanca de negoci: Element de decisió sobre el que una empresa té el control i sobre el que es pot realitzar una activitat comercial, ja sigui publicitària o retributiva per tal d'incrementar la rendibilitat.

Machine Learning: Tècnica informàtica que per mitjà de la repetició i la comparativa vers el resultat esperat, permet obtenir de manera automàtica un programa que millora en base a l'experiència sense estar explícitament programat.

Model predictiu: Branca dels algoritmes de Machine learning que estableix com a objectiu final una predicció numèrica o categòrica amb un % de confiança sobre un conjunt de dades.

Overfitting: Problema habitual en algoritmes de Machine learning en que una implementació s'adapta al conjunt de dades d'entrenament de manera excessiva i erra en les prediccions fora d'aquest conjunt.

Underfitting: Problema habitual en algoritmes de Machine learning en que una implementació no ha disposat de suficients dades d'entrenament i erra tant en les prediccions d'entrenament com en el cas generalitzat.

Insight: Coneixement extret d'un anàlisi de dades.

Lead: Persona que mostra un interès per una companyia, ja sigui per interacció amb publicitat online o per una sol·licitud directa d'informació.

Job: En bases de dades d'Oracle, procés que s'executa de manera automàtica en base a una periodicitat definida.

2. Prefaci

Aquest treball sorgeix en col·laboració amb l'empresa SDG Group a partir de les pràctiques realitzades prèviament per part de l'estudiant en l'empresa amb l'objectiu de millorar les capacitats d'anàlisi de dades, adaptació a les necessitats dels clients i finalment desenvolupament d'un projecte de forma autònoma.

De cara a les explicacions que es vagin fent sobre les dades, cal remarcar que s'han anonimitzat els noms del client, les variables i els valors d'aquestes ja que al tractar-se d'un treball d'investigació, l'empresa no vol que es faci públic a quin dels clients pertanyen les dades. Aquest fet pot provocar que en algun moment hi hagi una redacció ambigua o genèrica però en cas de ser necessari, es disposa de les dades sense anonimitzar.

2.1. Origen del projecte

Un dels clients de SDG Group que disposa d'una extensa base de dades on registren informació sobre clients, productes, ventes i venedors es planteja la possibilitat de realitzar un anàlisi amb l'objectiu de detectar si un client ja registrat a la base dades realitzarà una nova compra. En endavant, ens referirem a quest tipus de client com a *client renovador.*

Per a fer-ho, es planteja realitzar un estudi sobre l'estat de l'art actual i les diferents ofertes de mercat en l'anàlisi predictiu de dades i finalment, realitzar una proba pilot a fi de determinar si s'escau, quina de les alternatives es podria dur a terme.

2.2. Motivació

La principal motivació darrere d'aquest projecte radica en el fet que avui en dia som en una època digital on la nostre informació és a tot arreu. Per aquest motiu, les empreses busquen maneres d'explotar les dades disponibles d'una persona per tal de personalitzar ofertes i/o campanyes publicitàries amb l'idea de maximitzar la possibilitat de vendre.

Inicialment, s'utilitzava l'estadística descriptiva per a respondre a la pregunta "que ha passat?" i amb els resultats obtinguts, intentar millorar.

Aquest treball correspon a la següent etapa de l'anàlisi de dades, l'analítica predictiva, que pretén respondre la pregunta "que podria passar?" per adaptar-se als possibles resultats.

Per exemple, si sabem que cada any plou al setembre (analítica descriptiva) podem intentar portar un paraigües cada dia de setembre per intentar que no ens agafi

desprevingut.

D'altre banda, si aconseguim trobar un algoritme que en funció de la velocitat del vent, la humitat i els dies consecutius de sol ens fes una predicció del temps que farà, podríem utilitzar-lo per dur el paraigües només aquells dies en que l'algoritme prediu pluja.

Anàlogament, una empresa pot calcular estocs en funció de les seves estadístiques prèvies o pot intentar afinar molt més en base a prediccions de mercat.

3. Introducció

La tècnica de Machine learning, com el seu nom indica, es la implementació d'algoritmes que "aprenen" en base a la prova i error fins que aconsegueixen desenvolupar la tasca requerida amb una fiabilitat definida pel programador.

En aquest cas, la tasca que es pretén aconseguir es una predicció sobre la possibilitat de renovació d'un client en base a les dades recopilades per l'empresa sobre el client i les seves compres prèvies.

3.1. Context

El departament de I+D en conjunt amb la secció de màrqueting es plantegen trobar una manera de reduir costos i produir campanyes publicitàries molt més específiques per tal d'aconseguir un major rendiment (entenent com a rendiment la quantitat de vendes en relació a la quantitat invertida en aquestes campanyes).

Per a fer-ho, decideix avançar en el seu ús de les dades i sol·licita un estudi dels diferents productes de mercat per tal de començar en l'anàlisi predictiu.

SDG, com a empresa dedicada en anàlisi de dades, compta amb la seva pròpia proposta en aquest sector, anomenada DataRobot, en que ofereixen un servei extens però amb un cost elevat ja que com a experts en la matèria, garanteixen que si les dades són prou bones, obtindran resultats en algun dels possibles àmbits d'estudi.

D'altre banda, hi ha la possibilitat d'obrir un projecte a d'altres proveïdors de tipus RFP (*Request for Proposal*) en la que una empresa externa a la operadora actual, pot realitzar una proposta de projecte, calendari i preus. Aquesta alternativa, que podria aparentar més econòmica ve amb el risc afegit d'involucrar una nova empresa i per tant tractar amb un nou proveïdor que es desconeix si complirà els objectius que es marquin.

Finalment, es planteja una tercera opció en la que es desenvolupi una prova pilot, amb un cas d'ús concret per tal d'invertir un cost més baix, veure si les dades són prou correctes com per a treure'n conclusions i si a partir de l'estudi, s'obtenen resultats positius, decidir afrontar un projecte dels descrits prèviament.

A fi de posar en perspectiva les diferents solucions del mercat actual, presentem el gràfic publicat per Gartner, Inc sobre analítica predictiva:

Fig. 3.1. Gràfic comparatiu d'habilitat d'execució vers integritat de la visió. Font: [1]

Com podem veure, el software proposat pel proveïdor actual és complert i reconegut però n'hi ha d'altres que a priori podrien presentar millors solucions.

Així doncs, avaluem comparativament les diferents propostes en la següent taula a fi de determinar l'opció a escollir per al client:

	Proba pilot	Proveïdor actual	Proveïdor extern
Valoració econòmica	18.000€	500.000€	100.000€
Factor de risc [a]	Baix	Baix	Elevat
Durada [b]	Curt termini	Llarg termini	Mig termini
Compromís [c]	Mig	Alt	Baix

Taula 3.1. Comparativa de les diferents propostes existents.

[a] s'avalua el factor de risc com a la facilitat d'adaptar un procés nou al model actual i per tant la dificultat de complir amb els *timings* proposats en projecte vers la possibilitat de desviacions degut a imprevistos.
[b] s'avalua l'abast del projecte i la immediatesa a l'hora de prendre decisions de negoci derivades de les conclusions obtingudes.
[c] s'avalua l'oferta de solucions de negoci en base a les conclusions obtingudes.

Proba pilot: es planteja en els diferents aspectes tenint en ment aquest TFG. Per tant, consistiria en un estudi realitzat per un consultor que ja es troba a l'empresa i que per tant coneix el model de dades i s'estima en 45 jornades (12 ECTS * 30 hores/ECTS / 8hores per jornada) satisfent el cost estàndard de consultoria de 400€ / jornada. No es comprometen resultats ja que es tracta d'un pilot. Cal remarcar que en cap moment s'ha fet la proposta real al client ja que es tracta d'un treball de recerca amb el consentiment de l'empresa per usar les seves dades.

Proveïdor actual: el proveïdor actual, SDG, proposa augmentar el seu rang de serveis actual, afegint els estudis predictius amb la condició de contractar-lo com a un servei continu de 2 anys de durada. Aquesta opció, tot i tenir un cost elevat, comporta tractar amb una empresa que ja coneix els models de l'empresa i que esta acostumada a tractar amb el client de manera que representa una opció més segura. A més a més, es compromet contractualment a oferir solucions de negoci en base als resultats obtinguts.

Proveïdor extern: la recerca de nous proveïdors, representa incloure una empresa que no coneix el model actual, representant un risc major en la adaptació dels seus serveis però un cost molt menor i un projecte a mig termini ja que es definirien casos d'ús i un cop satisfets, es donaria per tancat el projecte. En aquest cas, davant la desconeixença de les dades que disposa el client, aquests proveïdors no garanteixen obtenir resultats.

A la vista de les alternatives, es decideix realitzar aquest pilot en el cas d'ús de clients renovadors. S'utilitzarà les dades de l'empresa per a seleccionar un set de dades de clients que han estat compradors en l'empresa de forma repetida i es separarà les dades en dos blocs, un d'aprenentatge per a l'algoritme i un de proba per a validar la fiabilitat de l'algoritme obtingut.

Finalment, es planteja l'opció de realitzar proves amb l'algoritme final en un set de dades sense categoritzar per a avaluar-ne la resposta.

3.2. Objectius del projecte

El projecte començarà per investigar les possibilitats actuals per a la implementació de l'algoritme, s'analitzaran les dades per a la correcta categorització i us d'etiquetes, es definiran aquestes etiquetes, s'implementarà l'algoritme i s'estudiaran els resultats per a definir millores i adaptar el procés fins aconseguir un predictor eficaç.

3.2.1. Objectiu 1 – Estat de l'art

Com a punt de partida, requerirem analitzar l'estat actual en matèria de Machine Learning.

S'avaluaran els diferents algoritmes d'ús generalitzat, se'n tractarà de definir els aspectes que els fan millors en certes tasques i en base a aquesta informació, es determinarà quin serà l'algoritme a emprar per al cas d'ús concret del treball.

Alhora, es probable que les pròpies dades de que disposem ens limitin o ens condicionin a escollir usar un algoritme i no un altre.

3.2.2. Objectiu 2 – Anàlisi i tractament de dades

Per a poder implementar un algoritme útil, s'haurà de validar que les dades que s'utilitzin com a referencia per a entrenar-lo no es veuen influenciades per factors externs als d'estudi.

Tenint en compte que es pretén predir el comportament dels clients en base a la informació dels productes que van comprar anteriorment i del seu perfil comercial, si les dades mostressin que es van disparar les vendes entorn a una campanya comercial, les variables dels clients que hi van participar estarien desviant el comportament real dels clients.

Similarment, per a poder reduir les variables que intervindran en l'estudi, reduint l'espai de combinacions possibles, es descartaran aquelles que resultin indiferents en base a l'estadística descriptiva i als criteris que la pròpia empresa determina en base a la seva experiència comercial.

Durant aquesta fase, s'emprarà el llenguatge SQL per a realitzar consultes a la base de dades i l'aplicació web SAP BO per a l'elaboració d'informes i la visualització de les dades d'una forma gràfica.

Finalment s'elaborarà un script que permeti automatitzar el tractament dels data set i la extracció dels mateixos per a les simulacions de l'algoritme.

3.2.3. Objectiu 3 – Disseny i implementació

En aquesta fase es realitzarà la implementació en si del programa.

Primer, es farà una comparativa dels mètodes més habituals a l'hora de desenvolupar aquest tipus d'algoritmes.

Seguidament, es dissenyarà la parametrització del algoritme escollit i es simularà el comportament amb els data sets preparats.

3.2.4. Objectiu 4 – Anàlisi de resultats

En qualsevol model de Machine learning, és pràcticament impossible obtenir el resultat més precís en el primer intent i en cas d'aconseguir-ho, es reduiria a una qüestió d'atzar.

Es per això que els algoritmes d'aquest tipus s'avaluen amb un procés iteratiu de millora fins a trobar una configuració en que la funció que s'utilitza per valorar la fidelitat del model no millora més.

Així doncs, l'última fase del projecte consistirà en realitzar diferents models, iterant en funció dels resultats obtinguts per assolir el millor resultat possible dins de l'abast del projecte.

3.3. Abast del projecte

A nivell funcional, es pretén crear un script, executable sense programari adicional, que emprant un model de Machine learning, sigui capaç de predir la probabilitat d'un client de tornar a comprar a l'empresa en base a característiques que el defineixen com a client.

La intenció de la metodologia triada es permetre a qualsevol usuari de l'empresa, independentment de les seves habilitats informàtiques, utilitzar el programa i realitzar prediccions.

Queda doncs fora de l'abast la decisió final del client sobre futurs estudis en base als resultats obtinguts.

A nivell de planificació, es proposa el següent calendari per a la realització dels diferents objectius definits prèviament per tal d'assegurar-ne el compliment.

Mes	Febrer			Març				Abril				Maig					Juny				Juliol					Agost				
Setmana	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	Jornades
Objectiu 1	1	1	1	2	1	1																								7
Objectiu 2						1	1	1	1	1																				5
Objectiu 3										1	1	1	1																	4
Objectiu 4													1	1	1	2	1	1	1	1										9
Objectiu 5																				1	1	1	1	1						5
Memoria			1		1			1			1				1			1					1	1	1	2	2	2		15
																													Total:	45

Fig. 3.2. Calendari proposta inicial de desenvolupament del projecte.

Dins l'apartat de redacció de la memòria s'inclou el temps invertit en la confecció de la presentació del treball. En el cas d'adaptar-se per a definir un projecte per al client, se substituiria la memòria per la documentació IT-PEP requerida (documentació del projecte, proposta, documentació funcional i manuals d'usuari).

4. Estat de l'art

Un cop definit el projecte, cal explorar la situació actual en la matèria, investigant i definint de forma clara les possibilitats que tenim al nostre abast.

Començarem doncs explorant els tipus d'algoritmes que s'utilitza actualment a la industria i recercant informació sobre els principals punts a tenir en compte a l'hora de triar-ne un.

4.1. Tipus d'algoritmes:

Dins del Machine Learning trobem 3 tipus generals d'algoritmes, que els categoritzen en base a les dades requerides i l'objectiu que cada tipus pretén assolir.

4.1.1. Aprenentatge supervisat:

Consisteix en definir de manera automàtica una funció que partint d'un conjunt de variables independents, calculi el resultat a obtenir.

Per al seu correcte funcionament, es requereix un conjunt de dades amb el resultat ja calculat a més de les dades per a les que volem fer servir l'algoritme. Aquest conjunt de dades del que ja coneixem el resultat servirà com a conjunt d'entrenament i serà el que l'algoritme utilitzarà per aprendre i idealment realitzar prediccions acurades.

Afegim el següent diagrama conceptual a mode d'exemple:

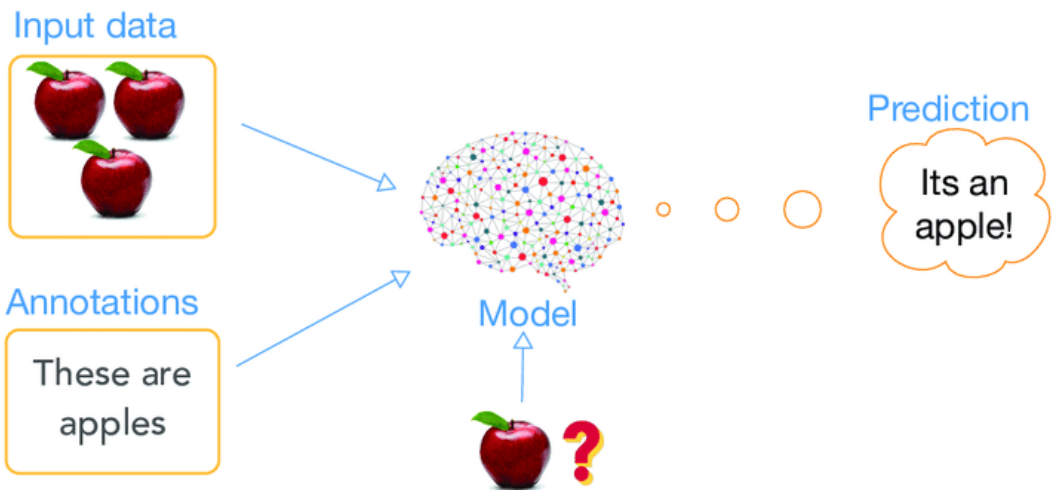


Fig. 4.1. Diagrama de funcionament d'un algoritme d'aprenentatge supervisat. Font: [2]

Alguns dels exemples d'aquest tipus d'algoritmes son la regressió, els arbres de decisions, els boscos aleatoris, anàlisi per K-pròxims-veïns (KNN) o la regressió logística entre altres.

4.1.2. Aprenentatge sense supervisió:

Es defineix com l'adaptació informàtica de l'aprenentatge Hebbià. Aquesta teoria d'aprenentatge, amb origen en l'estudi de les neurones, sosté que quan una cèl·lula activa un altre, la seva unió s'intensifica.

En l'adaptació, s'implementa com una suma ponderada de les variables d'entrada en que la ponderació es va corregint d'acord al resultat obtingut.

Les seves aplicacions consisteixen en la identificació de patrons prèviament desconeguts en un conjunt de dades sense categoritzar.

Així doncs, aquest tipus d'algoritmes busquen categoritzar o agrupar subconjunts de dades de forma automàtica per al seu posterior anàlisi.

Alguns dels exemples d'aquest tipus son les xarxes neuronals, les agrupacions jeràrquiques, k-mitjanes, models de barreja o categorització basada en densitat espacial entre altres.

4.1.3. Aprenentatge reforçat:

Aquest tipus d'aprenentatge busca perfeccionar la presa de decisions d'un software per a maximitzar una recompensa acumulada.

S'entén com un simulador d'escenaris en que s'exploren les possibles accions a prendre per un operador en un estat del sistema i es calcula la situació en el nou estat assolit. Afegim el següent diagrama conceptual a mode d'exemple:

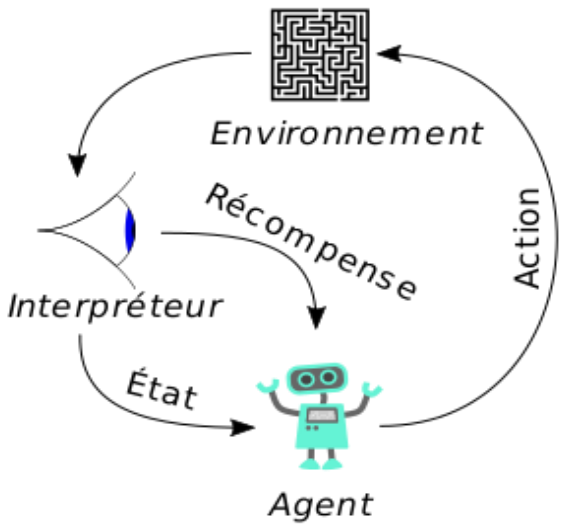


Fig. 4.2. Diagrama conceptual del funcionament d'un algoritme d'aprenentatge reforçat. Font: [3]

Alguns exemples d'aquest tipus son les simulacions de Monte Carlo, Q-learning o el procés de decisió Markovià.

4.2. Ampliació dels algoritmes:

Com el nostre cas contempla la decisió d'un resultat determinista per al qual disposem d'un conjunt de dades conegut i categoritzat, el nostre problema es resoldrà emprant un algoritme d'aprenentatge supervisat.

Explorem doncs els principals algoritmes d'aquesta categoria.

4.2.1. Regressió:

Un algoritme per regressió s'utilitza quan el resultat a calcular es un valor real o una variable continua, com podria ser un salari o el preu d'un producte.

La regressió lineal pretén trobar l'hiperplà que millor s'ajusti als punts que disposem com a conjunt de dades d'entrenament.

Pros: poca complexitat, fàcil d'entendre, fàcil d'interpretar, eficient en temps de computació.

Contres: totes les variables han de ser numèriques, només permet trobar relacions entre variables lineals e independents, es molt fàcil patir *overfitting*, es a dir adaptar-se de manera excessiva al conjunt de dades d'entrenament, fent que l'extrapolació a les dades d'anàlisi quedi completament esbiaixada; es poc robust vers dades anòmales o d'extrem.

4.2.2. Arbre de decisions:

L'arbre de decisions consisteix en modelar les dades com una presa de decisions seqüencials que determinen un resultat.

Generalment es representa mitjançant un graf en que els nodes son estats (valors concrets de les diferents variables) i les arestes els possibles canvis en aquests valors.

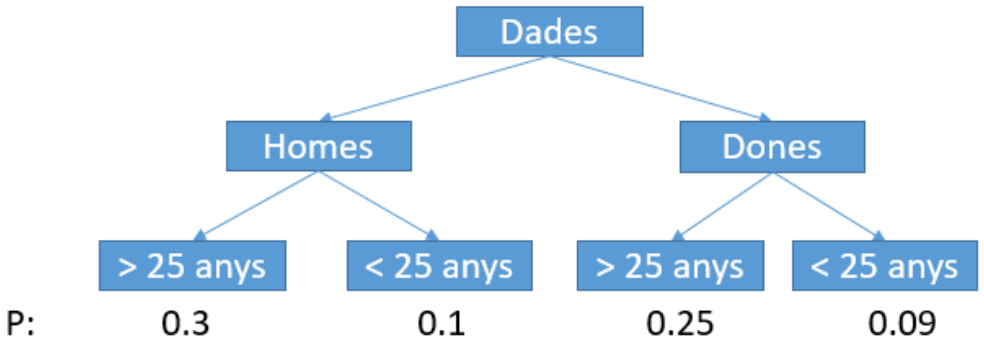


Fig. 4.3. Exemple d'elaboració pròpia per il·lustrar un arbre de decisions.

En aquest exemple, la probabilitat final retornada per l'algoritme es la de l'esdeveniment que intentem predir. Per exemple, per a la compra d'un producte seria la probabilitat de cada

subgrup d'efectivament comprar (es a dir, no han de sumar 100% amb els altres grups).

Aquests algoritmes es construeixen en dues etapes:

1 - Inducció: consisteix a dividir els valors possibles de les variables en subgrups de manera que, idealment, no hi hagi dues configuracions iguals en dos subgrups diferents.

2 - Poda: consisteix a millorar els temps de computació i la precisió dels resultats tallant les branques que utilitzen variables de poca rellevància. El mètode més senzill es coneix com a *reduced error pruning* i consisteix a eliminar branques que continguin poca informació. Per exemple, si tenim una branca que divideix un % petit de la mostra en 4 apartats per combinació de 2 variables en que 3 resultats són "SI" i 1 es "NO", es converteix en un node que indica directament aquest % com a "SI". Finalment, es comprova la precisió general del algoritme i es manté el canvi si no s'ha empitjorat.

4.2.3. Bosc aleatori:

Tal com el nom indica, un bosc aleatori (random forest) es basa en generar multitud d'arbres de decisions i donar com a resultat la moda dels resultats individuals en el cas d'un algoritme amb variables discretes o la mitja aritmètica dels resultats individuals en cas de treballar amb variables continues.

Les primeres implementacions d'aquest algoritme anomenades "bagging" es basaven en dividir el conjunt de dades d'entrenament en grups de mida aleatòria i entrenant un arbre de decisions a cada subconjunt.

Posteriorment, Ho, Tin Kam en el llibre "A Data Complexity Analysis of Comparative Advantages of Decision Forest Constructors" va proposar modificar l'algoritme per tal que els arbres individuals decidissin realitzar les divisions (branques) dels seus models de forma aleatòria enlloc de computant les divisions òptimes dels seus conjunts de dades.

L'argument es que si un model té una variable que influencia fortament el resultat, computant les branques òptimes dels subconjunts, molts dels arbres tendeixen a patir biaix per aquesta variable i acaben influenciant la decisió del grup.

Aquest model de divisions aleatòries, es el que es va anomenar "*random forest*".

4.2.4. K-propers-veïns (KNN):

Aquest algoritme planteja l'assignació de resultats en base a la proximitat a altres resultats coneguts (conjunt d'entrenament) formant clústers.

La resposta del model es basarà en assignar a un punt com a valor, la moda en cas de variables discretes i la mitja aritmètica en cas de variables continues, dels K punts més propers al d'anàlisi.

Els paràmetres d'aquest model són la funció emprada com a distancia entre punts i la variable K que determinarà la distancia "frontera" tal que el nombre de punts coneguts dins de la frontera = K.

Els punts dins d'aquesta frontera són els que participaran en la decisió del valor del punt d'anàlisi.

Plantegem un exemple per a entendre millor aquest algoritme:

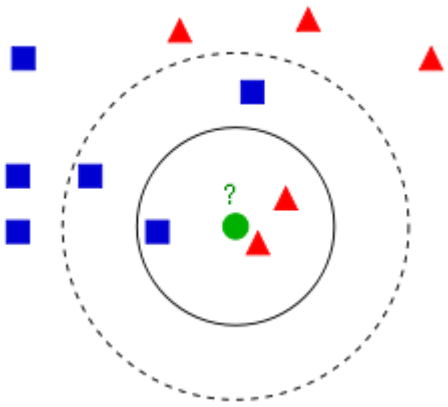


Fig. 4.4. Gràfic il·lustratiu del funcionament d'un algoritme de KNN

Intentem determinar si el cercle verd correspon a un triangle o a un quadrat.

Usem la distancia euclidiana i observem que, si determinem k = 3 (cercle sòlid) el vot definiria el punt com a triangle (per 2 vots contra 1 de quadrat) mentre que, si determinem k = 5 (cercle discontinu) quedaria definit com a quadrat (per 3 vots contra 2 de triangles).

Comprovant la precisió amb el conjunt de dades de validació podrem determinar quin valor de K es el més adequat per a la tasca a resoldre.

4.3. Elecció de l'algoritme

Tot i determinar la tipologia del nostre problema, caldrà definir de forma objectiva el motiu pel qual utilitzarem un algoritme i no un altre dels que hem explicat al apartat anterior.

Per a fer-ho, utilitzarem la comparació dels següents paràmetres en el nostre problema i triarem aquell que en el nostre cas obtingui un millor resultat en la majoria.

4.3.1. Parcialitat vers variància:

Suposem que disposem de diversos conjunts d'entrenament igual de bons. Es diu que un predictor pateix parcialitat quan per a un input X, independentment del conjunt d'entrenament, realitza de forma sistemàtica prediccions incorrectes per aquest X.

Paral·lelament, definim la variància del predictor com la quantitat de prediccions diferents que proporciona per al mateix input X en base al conjunt d'entrenament emprat. Un alta variància, per tant, impossibilitaria obtenir resultats fiables a l'hora d'usar el predictor.

Generalment, hi ha un equilibri entre parcialitat i variància ja que un algoritme ha de ser flexible en quant a les prediccions per adaptar-se a les dades d'entrenament però si esdevé massa flexible, patirà alta variància.

4.3.2. Complexitat de la funció i dades d'entrenament:

El segon aspecte a tenir en compte alhora de triar l'algoritme a emprar es avaluar el volum de dades necessàries per a entrenar-lo i les dades disponibles.

Si la funció (que pot no existir com a tal) que determina els resultats reals a partir de les variables d'estudi es simple, un algoritme rígid amb alta parcialitat podrà determinar-la amb un conjunt de dades petit.

D'altre banda, si la funció es complexa (per exemple perquè depèn de la interacció de diverses variables i es comporta de maneres diferents per a diferents valors d'estudi), aleshores caldrà un gran volum de dades d'entrenament i un algoritme flexible amb alta variància per a realitzar prediccions correctes.

4.3.3. Dimensionalitat:

En tercer lloc, es important entendre que si un problema parteix d'alta Dimensionalitat (gran quantitat de variables a partir de les que realitzar una predicció) la funció a predir esdevindrà complexa encara que la funció real només depengui d'un petit subconjunt d'aquestes variables.

El problema radica en que les variables "extra" confonen l'algoritme i provoquen alta variància. Per tant, per a corregir-ho caldrà modificar el predictor per a que utilitzi major parcialitat i podríem caure en el problema definit al punt previ.

Així doncs, sempre que sigui possible, caldrà eliminar les variables que resultin irrellevants.

4.3.4. Soroll:

Un altre aspecte a tenir en compte, típic en qualsevol anàlisi de dades, es el soroll.

Si el resultat desitjat es sovint incorrecte (per exemple per error humà o per errors en la lectura de sensors) aleshores potser l'algoritme no hauria d'intentar determinar la funció que dona el mateix resultat exacte que en el conjunt d'entrenament.

Si intentem que el predictor s'adapti a totes les possibilitats generades pel soroll, acabarem tornant a caure en el problema d' *overfitting*.

De fet, es pot caure en errors de soroll sense que existeixin errors de mesura (conegut com a soroll estocàstic) si la funció que intentem determinar es massa complexa per al model emprat. En aquest cas, la part de la funció que no pot ser modelada "corromp" les dades i provoca el que es coneix com a soroll determinista.

En cas de tenir qualsevol dels dos tipus de soroll en el problema plantejat, es millor utilitzar un predictor amb major parcialitat i menor variància.

4.3.5. Altres factors rellevants:

Heterogeneïtat de les dades: generalment, s'ha de tenir en compte el tipus de dades tractades. Mentre que alguns algoritmes com els arbres de decisions, son particularment robusts en aquest aspecte, molts altres requereixen que tots els imputs siguin numèrics i estiguin escalats a un rang similar de valors (p.e. [-1 , 1]).

Redundància de dades: si els paràmetres que utilitzem com a imputs del sistema presenten correlacions entre variables o altres redundàncies, podem trobar-nos amb que alguns algoritmes, particularment els que treballen amb distàncies, donin un resultat pobre degut a inestabilitats numèriques. Sovint, aquest problema es resol amb algun tipus de regularització.

Presencia d'interaccions i no-linealitat: si totes les variables tenen una aportació independent al resultat, els algoritmes basats en funcions lineals i de distància funcionaran de manera òptima però si hi ha interaccions complexes, aleshores els algoritmes com les xarxes neuronals i els arbres de decisions treballaran molt millor, en gran part perquè han estat especialment dissenyats per a trobar aquest tipus d'interaccions.

Validació creuada: finalment, al treballar amb algoritmes de Machine learning, la manera típica de veure quin algoritme o quina parametrització de l'algoritme obté millors resultats és experimentalment. Per a avaluar i comparar, s'utilitza la validació creuada, que consisteix en separar aleatòriament la mostra de dades en particions diferents per als conjunts d'entrenament i avaluació, calcular la mitjana aritmètica dels resultats obtinguts i comparar els models. D'aquesta manera assegurem que els resultats de l'algoritme són independents de la partició emprada.

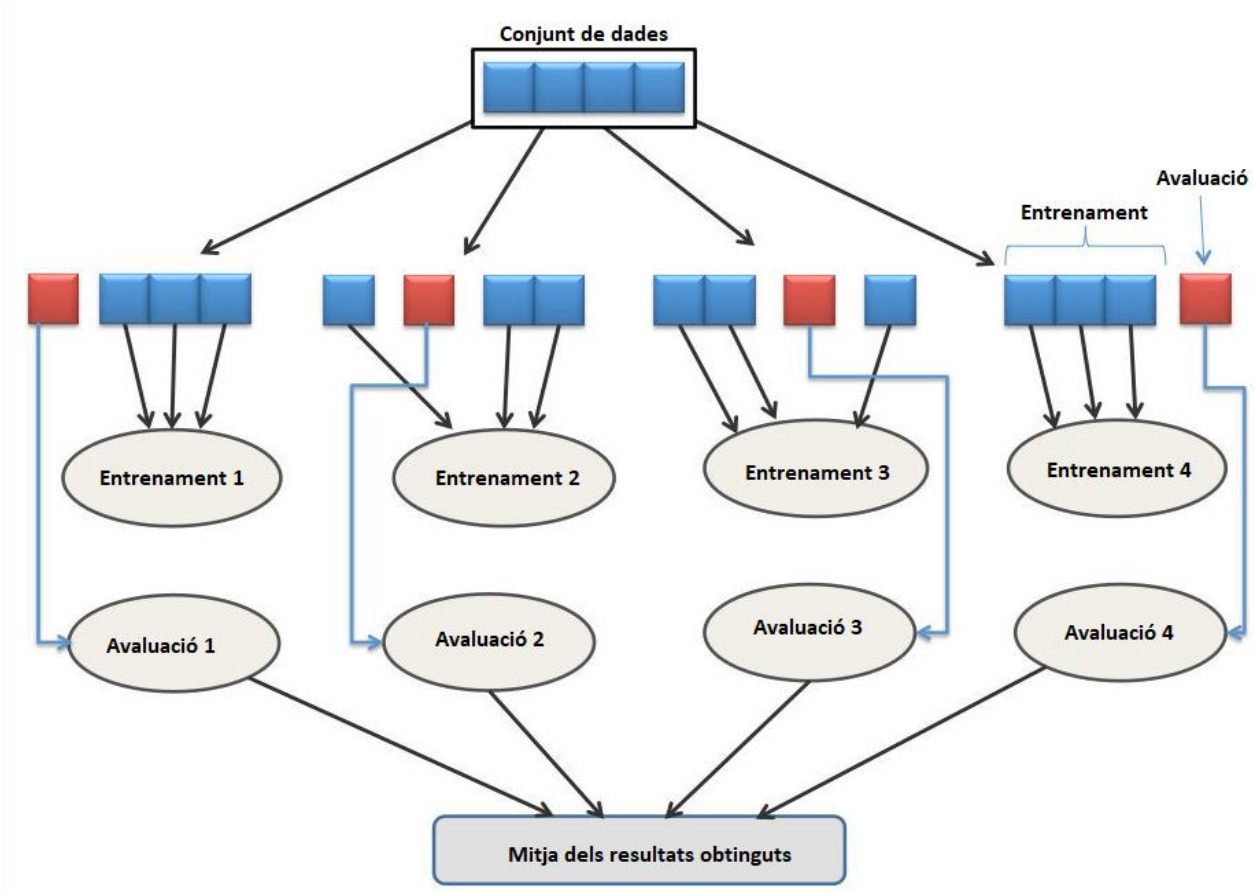


Fig. 4.5. Gràfic il·lustratiu del funcionament de la validació creuada.

4.3.6. Comparació d'algoritmes:

Veiem per tant, la comparació dels algoritmes proposats com a solució en la seva resposta davant els aspectes a tenir en compte descrits prèviament.

	Regressió	Arbre de decisions	Bosc Aleatori	KNN
Parcialitat	Rígid	Flexible	Molt Flexible	Flexible
Complexitat	Poques dades	Moltes dades	Moltes dades	Neutre
Dimensionalitat	Mala resposta	Bona resposta	Molt bona resposta	Molt bona resposta
Soroll	Bona resposta	Tendència overfit	Neutre	Bona resposta
Heterogeneïtat	Requereix tractament addicional	Robust	Molt robust	Robust
Redundància	Requereix regularització	Robust	Robust	Robust
Interaccions	Mala resposta	Bona resposta	Molt bona resposta	Mala resposta

Taula 4.1. Comparativa dels diversos algoritmes de major utilització en base als paràmetres descrits prèviament.

Tenint en compte que en el nostre cas tenim un problema amb moltes variables, interaccions que desconeixem, dades completament heterogènies i disposem d'un volum elevat de dades*, concloem que el model que millor s'adequa es el Bosc Aleatori.

* Basat en la norma "un de cada 10" que aproxima un valor mínim de mostreig com **10 * K / p** on **K** es el nombre de variables d'estudi i p la probabilitat de l'esdeveniment que volem predir.

5. Anàlisi i tractament de dades

Per a poder realitzar un estudi d'analítica avançada, primer haurem d'estandarditzar els possibles valors de les variables discretes, mostrejar les dades disponibles, establir els paràmetres a tenir en compte i finalment realitzar un script que permeti al usuari extreure el data set.

5.1. Estandardització

Tal i com hem anat esmenant, els algoritmes predictius son capaços de treballar amb variables discretes però son molt sensibles als canvis i al volum de valors possibles (ja que un valor excessivament elevat de variables podria elevar el temps de computació de forma dràstica).

5.1.1. Neteja de dades

Inicialment, emprem funcions conegudes com *regular expressions* per a eliminar accentuacions i símbols i convertim el text a majúscules.

Això permet no crear distincions entre variables que realment representen el mateix, com podria ser el municipi de "MALAGA" escrit en diferents fonts d'informació com a "Màlaga", "Málaga", "Malaga" o el propi "MALAGA".

Adjuntem el script de neteja a l'apartat de l'annex "Exemple de neteja de dades".

Fet això, veiem que aquesta primera fase de regularització no permet resoldre en tots els casos la unificació de valors que volen dir el mateix. Per exemple, la presencia d'informació en més d'un llenguatge suposa un altre problema ja que en aquests casos, no es tractarà d'una substitució simple de caràcters sinó d'una traducció. Similarment, tampoc es podrien resoldre a priori casos en que la font d'informació proporciona dades amb errors tipogràfics o ortogràfics.

Així doncs, es planteja emprar addicionalment un model d'estandardització a partir de diccionaris per aquelles variables que disposen d'un valor que volem establir com a "correcte".

5.1.2. Diccionaris

Els diccionaris son taules de valors en que la primera columna es el valor "correcte" que volem permetre a la variable, la segona columna es el mateix valor, per aquells casos en que la informació origen ja es correcte i les següents columnes contindran tots els valors possibles que trobem de variacions per a la mateixa variable.

Posem com a exemple el diccionari de províncies, que estandarditza valors amb errors tipogràfics i ortogràfics, a més de traduccions, al seu valor correcte:

	column1	column2	column3	column4	column5	column6	column7
9	A CORUÑA	A CORUÑA	LA CORUÑA	CORUNNA	A CORUNNA	15	15
10	ALAVA	ALAVA	VI	ARABA	ALABA	1	1
11	ALBACETE	ALBACETE	ALBACETE	AB	ALVACETE	2	2
12	ALICANTE	ALICANTE	A	ALACAN	ALACANT	3	3
13	ALMERIA	ALMERIA	ALMERÍA	ALMERIA	AL	4	4
14	ASTURIAS	ASTURIAS	ASTURIAS	O	ASTURIES	33	33
15	AVILA	AVILA	AV	HAVILA	ABILA	5	5
16	BADAJOZ	BADAJOZ	VADAJOD	VADAJOZ	BA	6	6
17	BARCELONA	BARCELONA	BARSELONA	B	BCN	8	8
18	BIZKAIA	BIZKAIA	VIZCAYA	VISCAIA	BI	48	48
19	BURGOS	BURGOS	BURGOS	BURGOS	BU	9	9
20	CACERES	CACERES	CÁCERES	CC	[NULL]	10	10
21	CADIZ	CADIZ	CADIS	CADIZ	CA	11	11
22	CANTABRIA	CANTABRIA	CANTABRIA	CANTABRIA	S	39	39
23	CASTELLO	CASTELLÓN	CASTELLÓ	CASTELLON	CS	12	12

Fig. 5.1. Exemple de diccionari extret de l'aplicació IDQ de PowerCenter.

La utilització d'aquests radica en unir la taula amb dades "origen" al diccionari per els camps 2 a N i retornar el valor de la columna 1 del diccionari en cas no-nul. Exemplifiquem la utilització a l'annex sota l'apartat "Exemple d'ús del diccionari".

Finalment, un cop satisfets amb els valors que prenen les variables d'estudi, passem a fer-ne un estudi exploratori per a detectar valors anòmals i les afectacions de cadascuna en l'output a determinar.

Es necessari aclarir que la part de diccionaris, s'ha inclòs a mode explicatiu per a justificar la correcta estandardització de les variables emprades però no s'ha desenvolupat com a part del treball ja que s'han generat i polit a la base de dades, amb el temps, per part de la nostre empresa com a tasca de manteniment perfectiu per a possibilitar estudis com aquest i d'altres.

5.2. Visualització de les dades

Per entendre millor la qualitat de les dades que disposem i posar en valor les afectacions dels diferents paràmetres per separat, s'ha realitzat una sèrie d'informes en l'aplicació SAP BO per a treballar de manera conjunta amb el client les variables que des d'un punt de vista de màrqueting haurien de tenir major influencia en determinar si un client tornarà a comprar un producte. Trobareu tots els scripts emprats per l'aplicació a l'annex.

Similarment, s'han estudiat variables que a priori no son determinants per a verificar-ho tot i que la decisió final d'incloure-les o no al model es definirà durant la implementació del algoritme.

Inicialment, es plantegen 2 blocs, l'enfocament al client i l'enfocament al producte.

5.2.1. Enfocament al client

Per un cantó, es planteja analitzar el comportament i la distribució dels clients en base a les variables que es disposa en quant a sexe, edat, província, freqüència d'utilització del producte i antiguitat com a client.

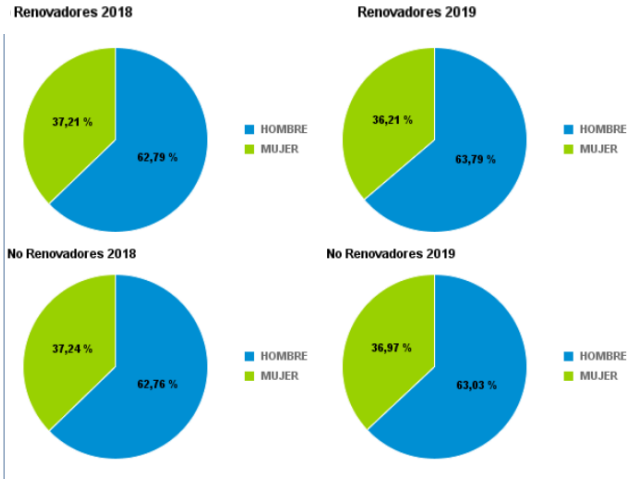


Fig. 5.2. Gràfic de distribució en la comparativa per sexe del client.

		Renovadores	Total	Ratio
2018	**Hombre**	3.701	30.346	12,20%
	Mujer	2.193	19.193	11,43%
2019	**Hombre**	2.792	25.904	10,78%
	Mujer	1.565	15.120	10,35%

Taula. 5.1. Taula de valors en la comparativa per sexe del client.

Veiem com en realitzar la comparativa, mentre que amb el canvi d'any hi ha una baixada d'un punt percentual en la renovació, tant la distribució de clients com les ratios de renovació es mantenen independentment del sexe del client.

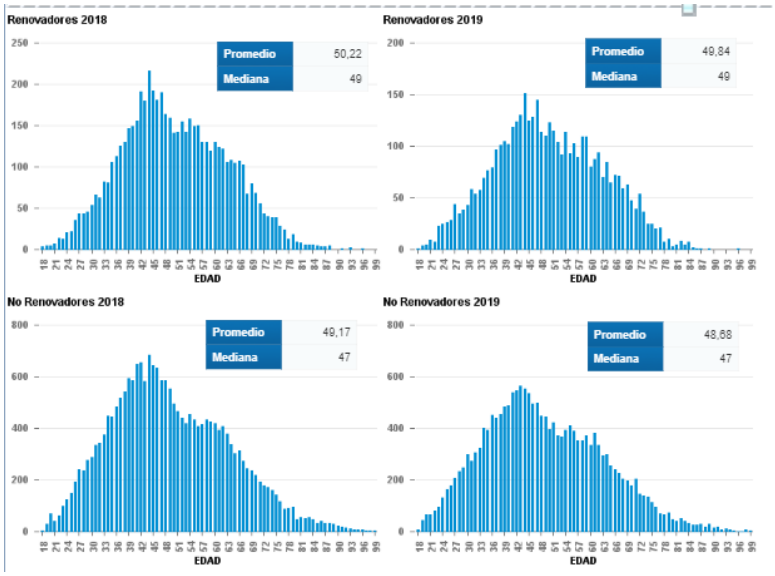


Fig. 5.3. Distribució de dades per edats dels clients.

Novament, veiem que segueixen una distribució normal entorn un valor mig però no es mostren diferencies significatives entre la distribució de clients renovadors vers la resta ni en les proporcions relatives.

DESC_PROV	Renovadores 2018	Renovadores 2019	Ratio 2018	Ratio 2019
MADRID	980	721	12,20%	10,85%
BARCELONA	864	724	10,50%	11,39%
VALENCIA	368	264	12,55%	11,38%
ALICANTE	269	203	11,82%	11,40%
CADIZ	248	211	13,59%	15,20%
SEVILLA	247	218	11,13%	12,42%
MALAGA	195	161	10,56%	10,77%
BIZKAIA	190	120	14,64%	11,41%
MURCIA	170	127	9,73%	9,08%
A CORUÑA	157	124	11,27%	10,61%
ASTURIAS	157	111	13,00%	11,96%
ILLES BALEA	156	101	23,25%	17,94%
GRANADA	145	104	13,03%	12,71%
TARRAGONA	121	95	10,48%	9,99%
GIRONA	120	91	10,79%	10,10%

Taula. 5.2. Comparativa de ratis de renovació per provincies. Es mostren els 15 primers valors en volumetria per simplicitat de visualització.

En aquest cas, sí que es mostren provincies amb tratis superiors a les demés pero no se'n infereix cap criteri lógic del estil "les provincies amb més població son mes propenses a

renovar" ni "les arees rurals renoven per sobre que les urbanes" cosa que impedeix establir un criteri per a l'empresa.

Horquilla Ki	Abs 2018	Ratio 2018	Abs 2019	Ratio 2019
0-50K	1.849	14,35%	1.583	14,02%
50K-100K	1.637	15,85%	1.072	13,41%
100K-150K	917	12,22%	627	10,27%
150K-200K	576	11,78%	402	10,54%
200K+	684	13,22%	526	12,93%

Taula. 5.3. Comparativa de renovació en base a la freqüencia d'utilització del producte previ.

Tanmateix, en aquest cas veiem que hi ha una forquilla que compren un rati superior als demés però no permet establir un control estricte sobre els clients ja que la utilització o no del producte no es quelcom controlable per l'empresa.

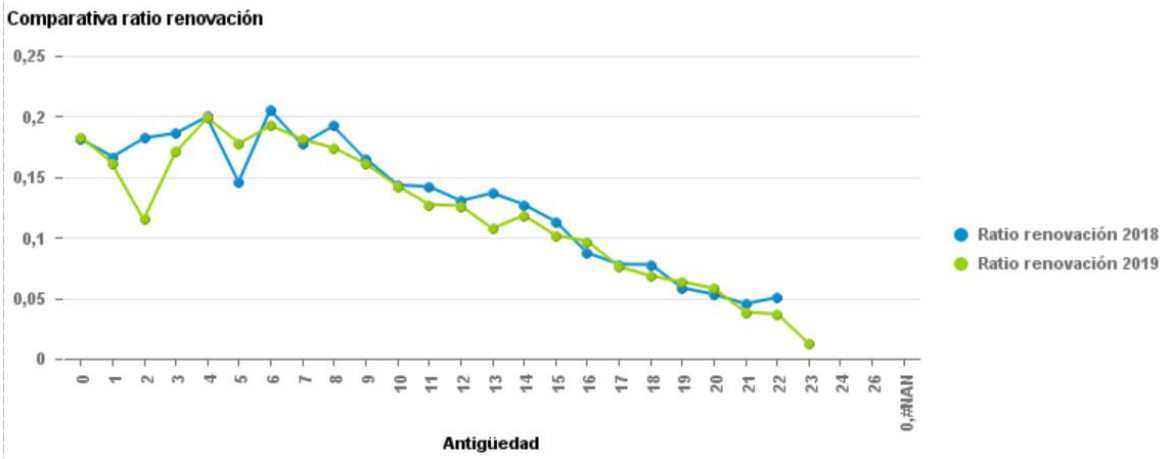


Fig. 5.4. Gràfic de ratio de clients renovadors en funció de l'antiguitat com a client.

Finalment es va analitzar el comportament dels clients en funció del temps, ja que es tenia la sospita que la ratio de clients que tornaven a confiar en l'empresa queia conforme passaven els anys.

Com es pot comprovar, aquesta sospita es confirma, cosa que esperem quedi reflectit en el model predictiu a l'hora d'evaluar clients de diferents antiguitats.

5.2.2. Enfocament al producte

Per l'altre cantó, l'empresa considera que probablement es tant important o més el producte que van comprar els clients com el propi perfil del client. En aquest cas, es considera que hi ha productes que poden haver triomfat més que d'altres i es planteja si aquest aspecte pot ser primordial a l'hora de decidir tornar a comprar o no.

Producte	Renovadores	Bajas 2018	Ratio renovac
Producte 1	13	74	17,57%
Producte 2	69	506	13,64%
Producte 3	51	759	6,72%
Producte 4	9	183	4,92%
Producte 5	0	45	0,00%
Producte 6	2.395	20.709	11,57%
Producte 7	54	448	12,05%
Producte 8	3	216	1,39%
Producte 9	822	8.487	9,69%
Producte 10	1	3	33,33%
Producte 11	1.252	13.058	9,59%
Producte 12	106	901	11,76%
Producte 13	33	325	10,15%
Producte 14	0	32	0,00%
Producte 15	784	3.412	22,98%
Producte 16	20	372	5,38%
Producte 17	401	2.788	14,38%
Producte 18	51	410	12,44%
Producte 19	22	107	20,56%
Producte 20	0	33	0,00%
Producte 21	56	187	29,95%
NO INFORMADO	2	14	14,29%

Producte	Renovadores	Bajas 2019	Ratio renovac
Producte 1	13	106	12,26%
Producte 2	24	387	6,20%
Producte 3	31	574	5,40%
Producte 4	16	151	10,60%
Producte 5	0	34	0,00%
Producte 6	1.849	16.360	11,30%
Producte 7	27	323	8,36%
Producte 8	7	157	4,46%
Producte 9	577	6.615	8,72%
Producte 10	1	4	25,00%
Producte 11	924	10.186	9,07%
Producte 12	67	675	9,93%
Producte 13	33	300	11,00%
Producte 14	0	66	0,00%
Producte 15	515	2.802	18,38%
Producte 16	25	370	6,76%
Producte 17	260	2.268	11,46%
Producte 18	92	666	13,81%
Producte 19	8	72	11,11%
Producte 20	0	20	0,00%
Producte 21	48	162	29,63%
Producte 22	0	89	0,00%
NO INFORMADO	5	27	18,52%

Taula. 5.4. Taula comparativa de ratis de renovació per producte donat de baixa.

Podem veure que efectivament hi ha productes que generen un rati de renovació superior als demés però la mostra es divideix de forma molt irregular i es podria arribar a pensar que molts casos son mal representats per la falta de valors. Davant això, l'empresa es planteja agupar els productes per categories (ja que alguns productes comparteixen característiques) per intentar trobar un patró explicable a nivell comercial. Un exemple seria poder obtenir un *insight* del estil "els productes de gama alta generen millor renovació que els de gama baixa" per a poder disposar d'una palanca de negoci.

Grup	Renovadores	Total 2018	Ratio 2018
Grup 1	2	14	14,29%
Grup 2	3.681	34.318	10,73%
Grup 3	855	4.194	20,39%
Grup 4	997	9.991	9,98%
Grup 5	434	3.145	13,80%
Grup 6	175	1.407	12,44%

Grup	Renovadores	Total 2019	Ratio 2019
Grup 1	5	27	18,52%
Grup 2	2.804	26.960	10,40%
Grup 3	632	3.927	16,09%
Grup 4	697	7.804	8,93%
Grup 5	293	2.634	11,12%
Grup 6	91	1.062	8,57%

Taula. 5.5. Taula comparativa del rati de renovació en relació a l'agrupació de productes.

Identifiquem el grup 3 com a destacable positiu i els grups 2 i 4 com a destacables negatius però observem que els altres grups es comporten de manera similar.

A la vista de tots aquests anàlisi l'empresa es planteja el següent: que passa en els casos en que un client compleixi un perfil comercial amb molt de potencial de renovació mentre que el seu producte es dels pitjors?

Com hem vist, totes aquestes variables son independents i per tant resulta complex decidir si unes prevalen sobre altres o si existeix alguna funció complexa que pugui determinar un resultat. Es per aquest motiu que es planteja l'objectiu d'aquest treball, verificar si un algoritme de Machine learning és capaç de trobar aquesta funció.

5.2.3. Altre informació disponible

Addicionalment a les variables descrites, la base de dades disposa d'altres aspectes que podrien influir en la decisió final del client. Tot i que el departament d'analistes tradicionals no les tenia en compte, s'han explorat i avaluat per a determinar si serien útils dins el model.

Respuesta	Clientes 2018	Renovadores 2018	Ratio
1	102	22	21,57%
2	81	28	34,57%
3	213	61	28,64%
4	828	259	31,28%
5	4.022	1.394	34,66%
6	35	10	28,57%
	47.464	4.369	9,20%

Respuesta	Clientes 2019	Renovadores 2019	Ratio
1	99	19	19,19%
2	66	14	21,21%
3	176	34	19,32%
4	763	173	22,67%
5	3.321	851	25,62%
6	31	9	29,03%
	37.623	3.455	9,18%

Taula. 5.6. Taula comparativa del rati de renovació vers la puntuació resposta en enquestes de satisfacció.

Veiem que els tamanys de mostra per a les enquestes de satisfacció son baixos i no mostren una tendencia ja que aparentment, tant si han contestat satisfets com si han contestat insatisfets, obtenen un rati de renovació de més del doble que els que directament no han respost mai cap enquesta.

Addicionalment, relacionar els clients directament amb les respostes que van donar no es una tasca simple i requereix de permisos addicionals per protecció de dades (normativa GDPR – perfilat de clients i cessió d'informació a 3rs ja que les enquestes no les realitza l'empresa de forma directa). Per aquest motiu, no s'inclourà al model.

Concepto	Renovadores	Ratio 2018
Acc1	1334	29,41%
Acc2	474	23,19%
Acc3	2606	19,08%
Acc4	454	17,80%
Acc5	989	17,84%
Acc6	123	16,06%
N/a	2103	7,18%

Concepto	Renovadores	Ratio 2019
Acc1	655	26,21%
Acc2	255	22,19%
Acc3	1236	17,42%
Acc4	252	17,52%
Acc5	483	16,58%
Acc6	52	13,79%
N/a	825	6,14%

Taula. 5.7. Taula comparativa del rati de renovació vers la compra d'accessoris juntament amb el producte.

L'empresa considera que els accessoris poden veure's influenciats per campanyes comercials i ventes creuades però disposen de dades i en analitzar-les, es manifesta que hi ha diferencia entre la gent que ha obtingut algún dels accessoris més significatius i addicionalment, n'hi ha que indiquen aspectes de la "personalitat" del client, cosa que creiem podria influenciar el resultat obtingut. En aquest cas, l'explotació d'aquestes dades no implica un retreball i per tant considerem que si es disposa de les dades, s'haruien d'incloure al model, com a mínim de forma inicial i si de cas eliminar-les en següents iteracions.

Flag Lead	Bajas 2018	Renovadores 2018	Ratio
0	51810	5833	11,26%
1	622	184	29,58%

Flag Lead	Bajas 2019	Renovadores 2019	Ratio
0	40295	4021	9,98%
1	1482	456	30,77%

Taula. 5.8. Taula comparativa del rati de renovació vers la categorització del client com a lead.

Aquesta variable es podria considerar part del perfil de client però s'ha tractat per separat ja que depen de molts altres factors. Un *lead*, es una persona que ha mostrat interés per l'empresa, fent click en anuncis online, demanant informació al web de l'empresa o contactant amb una de les botigues. Adicionalment, aquesta categoria té una durada limitada de 4 mesos i per tant resulta volàtil però com es por apreciar (i resulta evident dins de l'aspecte comercial del negoci), té un fort impacte en la ratio de renovació.

Horquilla	Renovadores	Ratio 2018
0	773	5,03%
1-5	5199	14,94%
5-10	867	14,84%
10-15	228	15,48%
15-20	71	17,93%
20-25	20	17,09%
+25	17	28,33%

Horquilla	Renovador	Ratio 2019
0	562	4,91%
1-5	3889	13,58%
5-10	636	13,33%
10-15	147	12,95%
15-20	61	18,21%
20-25	12	14,12%
+25	6	17,14%

Horquilla	Renovadores	Ratio 2018
0	1088	6,38%
0-5K	5333	14,83%
5-10K	342	15,01%
10-15K	71	17,44%
+15K	42	21,76%

Horquilla	Renovador	Ratio 2019
0	798	6,22%
0-5K	3985	13,49%
5-10K	244	13,22%
10-15K	39	12,70%
+15K	28	19,31%

Taula. 5.9. Taules comparatives del rati de renovació vers el nombre visites i despesa acumulada en milers € als tallers oficials.

L'empresa disposa de servei d'atenció als clients per a manteniment i reparació dels productes. S'ha analitzat doncs el nombre de visites dels clients i la despesa acumulada al llarg de la vida útil del producte que van estar utilitzant.

Veiem que el comportament en totes les franjes es similar excepte els 0. En veure les dades, el que es conclou es que els 0's son clients que realitzen les visites en serveis no oficials i per això no es disposa de les dades. Canvia doncs el criteri i s'utilitzarà al model una variable binària que indicarà si el client va utilitzar o no un servei oficial ja que aquesta variable si que influeix en el resultat.

Categoria	Clientes 2018	Renovadores 2018	Ratio
Sin expediente	46932	4823	10,28%
Categoria 1	1516	423	27,90%
Categoria 2	1074	138	12,85%
Categoria 3	743	136	18,30%
Categoria 4	500	125	25,00%
Categoria 5	474	83	17,51%
Categoria 6	331	112	33,84%
Categoria 7	293	49	16,72%
Categoria 8	237	87	36,71%
Categoria 9	199	33	16,58%
Categoria 10	192	34	17,71%
Categoria 11	186	62	33,33%
Categoria 12	185	70	37,84%
Categoria 13	164	30	18,29%
Categoria 14	154	16	10,39%
Categoria 15	148	72	48,65%

Categoria	Clientes 2019	Renovadores 2019	Ratio
Sin expediente	37366	3717	9,95%
Categoria 1	1287	261	20,28%
Categoria 2	869	96	11,05%
Categoria 3	600	89	14,83%
Categoria 4	374	59	15,78%
Categoria 5	358	51	14,25%
Categoria 6	245	62	25,31%
Categoria 7	224	37	16,52%
Categoria 8	177	46	25,99%
Categoria 9	170	34	20,00%
Categoria 10	159	21	13,21%
Categoria 11	152	43	28,29%
Categoria 13	149	23	15,44%
Categoria 15	134	44	32,84%
Categoria 14	130	8	6,15%
Categoria 12	117	17	14,53%

Taula. 5.10. Taula comparativa del rati de renovació vers les dades d'expedients i reclamacions en el servei d'atenció telefonic.

Per acabar d'exhaurir les dades disponibles, s'analitza el comportament dels clients en funció de les reclamacions en el *call center* de l'empresa. Aquestes trucades d'atenció al client es troben separades per categories (reclamacions, informació, queixes, etc.) però novament ens trobem amb que el comportament general es similar (tret de casos en que la mostra no es realment significativa) i per contra, sí que s'observa diferencia notable entre els clients que s'han posat en contacte alguna vegada vers els que no.

Així doncs, en aquest cas també s'emprarà un valor binari per indicar si hi ha algun expedient del client o no.

5.2.4. Eliminació de valors anòmals

Durant l'anàlisi exploratori de les dades hem detectat que en molts casos es detecta l'existència de valors que alterarien els resultats obtinguts. Per exemple, quan analitzem l'edat dels clients, ens trobem amb gent que apareix a la base de dades amb 999 anys o sense una data de naixement informada, comptabilitzant com a 0. Així doncs, en l'extracció del data set que servirà per alimentar el predictor, emprarem restriccions de filtrat usant la clàusula de SQL "WHERE".

```
WHERE EDAD IS NOT NULL AND
      EDAD <= 90
```

En aquest cas, d'aquesta manera assegurarem que ni el valor de l'edat es nul ni conté un valor enorme causa d'una errata en la digitalització o d'un selector en un formulari online.

En els altres casos s'han aplicat restriccions similars, fàcilment intel·ligibles al codi que s'adjunta a la memòria però s'ha emprat aquest exemple simplificat a mode il·lustratiu de la metodologia usada.

5.3. Extracció del data set

Per al punt final del tractament de les dades en sí, s'emprarà la funció del programa SAP BO d'extracció d'Excels a partir d'una taula muntada en la base de dades Oracle de l'empresa amb un procés PL/SQL.

5.3.1. Oracle SQL

Per a aquesta secció s'ha emprat el programa TOAD per a bases de dades d'Oracle ja que es l'eina utilitzada a l'empresa.

Adjuntem el script al script "p_dataset_monthly" adjunt a l'annex. Aquest script bàsicament genera dues taules de dades que contenen les que serviran per entrenar l'algoritme i les dades sobre les que volem realitzar prediccions.

En essència el que s'ha aconseguit amb els diferents estudis es generar un model "estrella" en que el centre es la taula mestra de clients i les diferents variables o atributs que s'han calculat en base a diferents actuacions dels mateixos (expedients, enquestes, etc.) s'adhereixen de manera modular. Així aconseguim una taula final on cada registre té les dades i valors del client i tots aquests atributs.

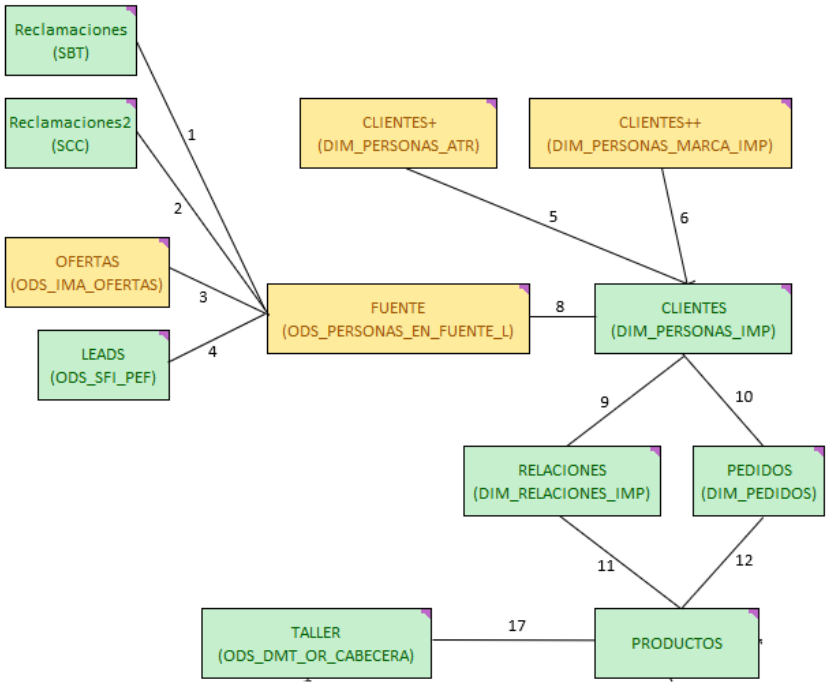


Fig. 5.5. Graf relacional de les taules de la base de dades utilitzada.

Conceptualment, aquest es el graf de les taules en base de dades emprades. Es disposa també d'un fitxer Excel en que es mostra el mapa complert i les anotacions pertinents als nombres que es veuen en els enllaços entre taules (indicant els camps per a fer JOIN).

De cara al client, aquesta part no forma part de l'entregable final (executable) ja que el procés s'automatitzarà per actualitzar les dades de forma mensual mitjançant un *job*, una configuració específica per a scripts en PL/SQL que permet indicar a la base de dades cada quant de temps ha d'auto-executar-se.

5.3.2. SAP BO

L'aplicació SAP BO permet configurar extraccions per a ser enviades de forma directa al correu de l'usuari, de manera que tampoc requereixi cap acció per part seva.

Això s'aconsegueix generant un informe a partir d'una query a la base de dades.

Aquest informes tenen la capacitat de generar gràfics i eines de reporting visuals per a que la gent de negoci rebi dades o status de l'empresa de forma àgil però també permet generar taules simples en format Excel per a extreure data sets de forma automàtica.

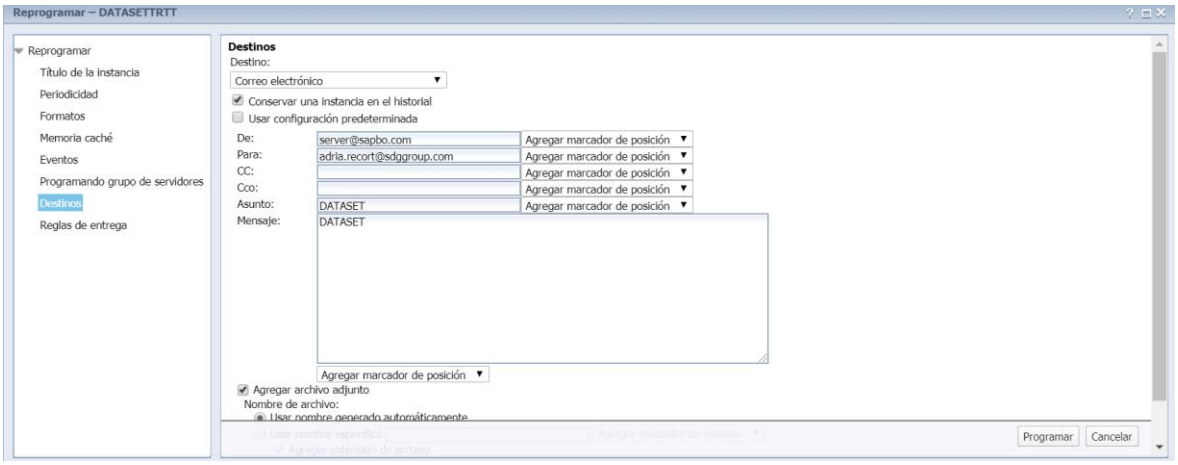


Fig. 5.6. Captura del aplicatiu SAP BO on es mostra la configuració del automatisme per correu electronic.

A l'apartat "periodicidad" es defineix un enviament mensual el dia i hora convinguts per l'empresa i a la secció "formatos" es defineix que s'enviarà l'extracció en Excel ja que d'aquesta manera el document obtingut serà llegible pel predictor sense necessitat de formateig.

El contingut i títol del missatge es modificable pero es mostra l'exemple que s'ha emprat com a demostració de funcionament.

Com podem veure, el servidor ens envia un correu d'acord a la programació:

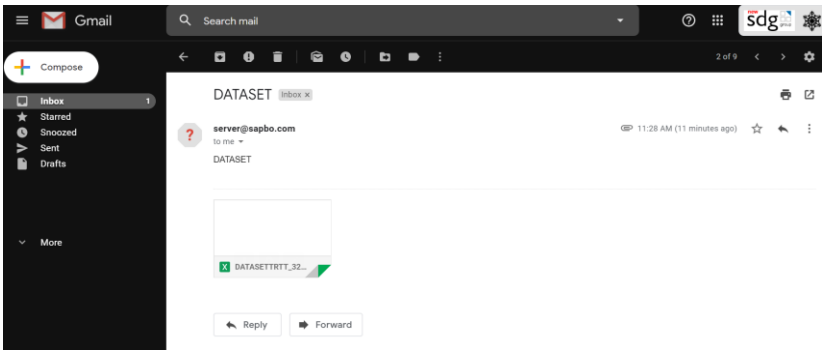
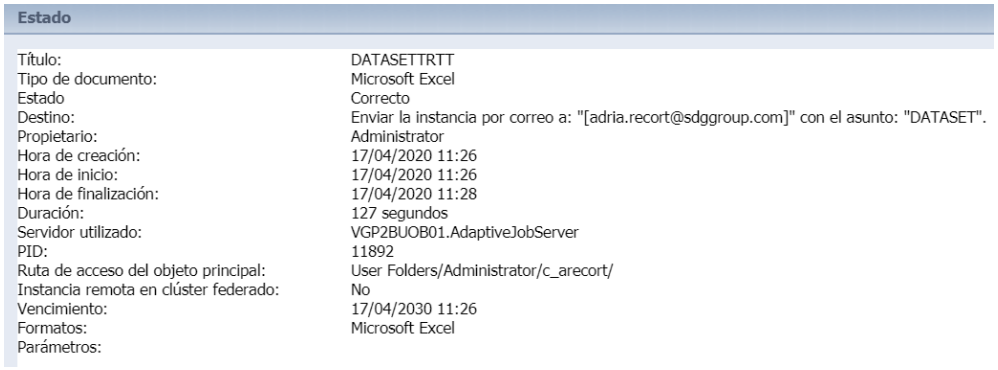


Fig. 5.7. Captura del correu de l'usuari receptor del dataset automatittzat.

Finalment el servidor ens mostra un missatge resum del estat de l'enviament com a comprovació per al futur manteniment del flux de dades.



Estado

Título:	DATASETTRTT
Tipo de documento:	Microsoft Excel
Estado	Correcto
Destino:	Enviar la instancia por correo a: "[adria.recort@sdggroup.com]" con el asunto: "DATASET".
Propietario:	Administrator
Hora de creación:	17/04/2020 11:26
Hora de inicio:	17/04/2020 11:26
Hora de finalización:	17/04/2020 11:28
Duración:	127 segundos
Servidor utilizado:	VGP2BUOB01.AdaptiveJobServer
PID:	11892
Ruta de acceso del objeto principal:	User Folders/Administrator/c_arecort/
Instancia remota en clúster federado:	No
Vencimiento:	17/04/2030 11:26
Formatos:	Microsoft Excel
Parámetros:	

Fig. 5.8. Captura del missatge registrat a la consola d'administració de l'aplicatiu per notificar el correcte enviament.

6. Disseny i implementació del predictor

Un cop disposem d'un conjunt de dades per a treballar-hi, es planteja el dubte de la implementació del algoritme en si.

Tot i haver definit al capítol 4 que utilitzaríem un Bosc Aleatori, en fer recerca, veiem que les implementacions varien substancialment en funció del propòsit, les dades i el nivell de detall de cara a parametritzar l'algoritme per a millorar-ne la precisió.

Addicionalment, pretendre implementar de manera autònoma l'algoritme, implicaria el risc que un error a la programació comportés una mala utilització de les dades, resultats incorrectes, etc.

Així doncs i d'acord amb l'abast del projecte s'ha optat per utilitzar la variant de l'algoritme de bosc aleatori de lliure utilització recomanat per la Dra. Alexis Cook a [4] anomenat XGBoost (*extreme gradient boosting*).

Aquest algoritme genera arbres de decisió de forma aleatòria i els va acumulant, donant com a output la mitja resultant dels diferents arbres, mentre l'error es va reduint. Si l'error augmenta, no s'afegeix al model i l'algoritme finalitza si s'arriba a un error mínim.

6.1. Parametrització

L'algoritme escollit presenta 2 paràmetres principals de definició del model i 2 paràmetres d'entrenament o definició de l'aplicació:

Nº estimadors: el nombre d'estimadors determina el nombre de vegades que l'algoritme realitzarà el bucle descrit (generar un model aleatori, afegir-lo als existents i calcular l'error de predicció). Un valor massa baix per aquest paràmetre comportaria una tendència a *underfitting* i per tant a resultats incorrectes tant en el set d'entrenament com al d'avaluació. D'altre banda, un valor massa elevat provocaria *overfitting* i per tant bons resultats al set d'entrenament però mals resultats en l'avaluació i futures prediccions.

Rati d'aprenentatge: aquest paràmetre actua com un valor de ponderació per al càlcul de la mitja resultat. Si fixem aquest en un valor petit, podrem elevar el nombre d'estimadors sense patir *overfit* però s'ha de tenir en compte que com més estimadors pretenguem utilitzar, més tardarà l'algoritme en entrenar el model.

Nº de tasques paral·leles: aquest paràmetre ens permet ajustar el nombre de tasques paral·leles que utilitzarà l'algoritme per a reduir el temps de computació. En conjunts de dades petits, aquest valor perd rellevància però per a conjunts de dades grans i amb moltes variables, resulta convenient ajustar aquest valor al nombre de nuclis del computador per a agilitzar els càlculs.

Nº de rondes per aturar: com hem descrit en la definició del algoritme, aquest s'aturarà arribat a un error mínim. Pot passar però que de forma aleatòria, l'algoritme trobés un arbre aleatori que s'adeqüés molt bé al set de dades d'entrenament, obtenint un model que realment no escalaria correctament a altres dades (avaluació i futures). Per tant, es assenyat determinar aquest valor a 5 per exemple, per a que si durant 5 iteracions el model no millora, doni el resultat per acabat. Aquest paràmetre, si s'especifica, sobreescriu al nombre d'estimadors definits inicialment. En cas de seguir millorant totes les iteracions o no definir-lo, el programa iterarà fins al nombre d'estimadors definits.

Cal mencionar que l'algoritme disposa de molts més paràmetres, disponibles a [8] que serveixen per a especificar condicions molt concretes per a l'optimització d'execució i càlcul que resulten complexes d'entendre i aplicar i queden fora de l'abast d'aquest projecte.

6.2. Entrenament, validació i predicció

Per a la part experimental de la implementació de l'algoritme, disposem de dos conjunts de dades, el conjunt d'entrenament i el conjunt de predicció que hem preparat amb el script de PL/SQL. Tot el codi en endavant, ha estat realitzat en python i es troba a l'annex.

6.2.1. Primer cas pràctic

Per al primer cas de proba del algoritme, dividim les dades d'entrenament de manera aleatòria usant l'eina *train_test_split* de la llibreria *sklearn*.

Parametritzem l'algoritme amb un nombre d'estimadors = 1.000, un rati d'aprenentatge de 0,05 i un nombre de rondes per aturar de 5 i obtenim en 20 segons, un resultat amb un error absolut mitjà de 0,2453

Observem que per a ser un primer intent, el programa ha resolt ràpidament i encerta aproximadament 3 de cada 4 clients mentre que recordem que el rati global que intentem predir ronda el 30% dels casos, es a dir, que una persona triant aleatòriament només encertaria 3 de cada 10 clients. Per tant, podem dir que aquesta primera aproximació resulta prou bona.

6.2.2. Iteració per paràmetres

Realitzem una segona prova per demostrar que en el cas dels predictors i del Machine learning, la frase “com més millor” no sempre s’adequa a la realitat.

Parametritzem l’algoritme amb nombre d’estimadors = 10.000 i un rati d’aprenentatge de 0,001 però passats 9 minuts, obtenim un error de 0,2462

Demostrat això, decidim iterar sobre els diferents paràmetres per a experimentar amb la resposta de l’algoritme i trobar, si existeix, un valor idoni de cadascun. Essencialment, implementarem el predictor sobre les mateixes dades i anirem enregistrant els valors que obtenim per a l’error en la predicció i el temps de computació de l’algoritme.

Començant amb el nombre d’estimadors en 1.000 incrementant en 500, obtenim el següent:

n_estimators	error	time
1000	0.24548969072164947	0:00:12.975357
1500	0.24548969072164947	0:00:12.819268
2000	0.24548969072164947	0:00:12.676256
2500	0.24548969072164947	0:00:12.668632
3000	0.24548969072164947	0:00:12.764419
3500	0.24548969072164947	0:00:12.701177
4000	0.24548969072164947	0:00:12.698596
4500	0.24548969072164947	0:00:12.874702
5000	0.24548969072164947	0:00:12.773167

Fig. 6.1. Resultats obtinguts en iterar sobre el nombre d’estimadors.

Veiem com clarament l’algoritme s’està aturant abans d’arribar als 1.000 estimadors ja que no pot millorar més. Provem doncs, en el cas del rati d’aprenentatge i mantenint 1.000 estimadors, començant en 0,05 i reduint el rati un 20% cada iteració i obtenim:

learning_rate	error	time
0.05	0.24548969072164947	0:00:12.274173
0.04	0.24604197349042708	0:00:09.973328
0.032	0.24714653902798234	0:00:12.143523
0.0256	0.24475331369661266	0:00:14.967970
0.0205	0.24714653902798234	0:00:18.190352
0.0164	0.24641016200294552	0:00:25.443952
0.0131	0.24659425625920472	0:00:31.860526
0.0105	0.24548969072164947	0:00:40.087521
0.0084	0.24447717231222385	0:00:46.108560
0.0067	0.24705449189985274	0:00:50.414547
0.0054	0.24558173784977907	0:01:09.827048

Fig. 6.2. Resultats obtinguts en iterar sobre el rati d’aprenentatge amb 1.000 estimadors.

En aquest cas, observem que el rati d'aprenentatge no permet a l'algoritme millorar en les prediccions. Per a verificar-ho realitzarem un últim assaig amb els estimadors assignats a 10.000 esperant que el temps de computació augmenti però els resultats no millorin.

learning_rate	error	time
0.05	0.24548969072164947	0:00:12.421415
0.04	0.24604197349042708	0:00:10.093356
0.032	0.24714653902798234	0:00:14.367146
0.0256	0.24475331369661266	0:00:20.006393
0.0205	0.24714653902798234	0:00:20.021875
0.0164	0.24641016200294552	0:00:27.536907
0.0131	0.24659425625920472	0:00:28.961856
0.0105	0.24548969072164947	0:00:36.116792
0.0084	0.24447717231222385	0:00:44.389342
0.0067	0.24705449189985274	0:00:48.270083
0.0054	0.24558173784977907	0:01:06.272053

Fig. 6.3. Resultats obtinguts en iterar sobre el rati d'aprenentatge amb 10.000 estimadors.

Tal i com esperàvem, els resultats obtinguts no milloren en qualitat però tampoc empitjoren en temps de computació. Això es deu a que encara que reduïm el rati d'aprenentatge, l'algoritme es segueix aturant en un valor inferior al màxim de predictors i per aquest motiu, encara que augmentem el màxim possible, els temps es mantenen.

Finalment, els paràmetres que utilitzarem en la implementació final seran 1.000 estimadors i un rati d'aprenentatge de 0,04 ja que han estat els valors que han obtingut millor temps de computació i resultat.

6.2.3. Validació creuada

Tal i com s'ha mencionat al apartat 4, addicionalment a la selecció correcte de paràmetres, la tècnica més utilitzada en el Machine learning per a validar el correcte funcionament d'un model consisteix a dividir el conjunt de dades disponible de diverses maneres i executar l'algoritme sobre les diferents permutacions de conjunt d'entrenament i conjunt de validació per a comprovar que efectivament el model que s'ha desenvolupat no depèn de com s'han repartit les dades.

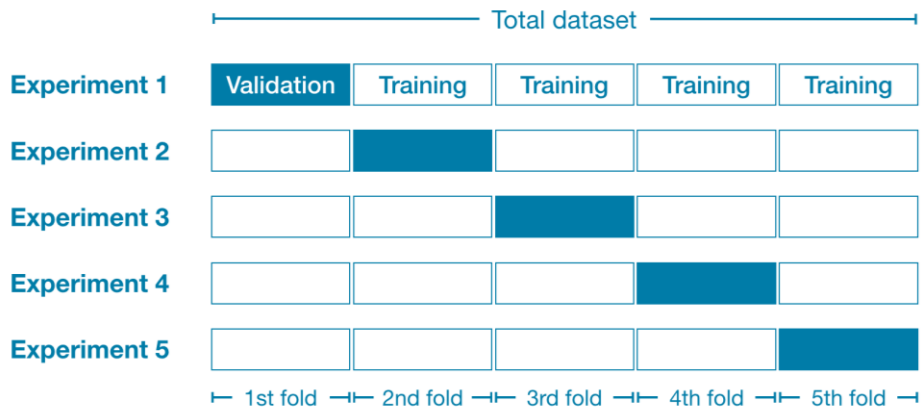


Fig. 6.4. Repartiment de les dades en una validació creuada mitjançant pipelines. Font: [4]

En aplicar aquest procediment, obtenim el següent:

```
from sklearn.model_selection import cross_val_score


scores = -1 * cross_val_score(my_pipeline, X_train,y_train,
                              cv=5,
                              scoring='neg_mean_absolute_error')

print("Average MAE score:", scores.mean())
```

```
Average MAE score: 0.24152556266624509
```

```
scores
```

```
array([0.23909792, 0.23403521, 0.24116903, 0.24303797, 0.25028769])
```

Fig. 6.5. Captura de Jupyter Notebook amb els resultats de la validació creuada

Validem doncs que el nostre predictor respon correctament a les dades utilitzades per a entrenar-lo independentment de les agrupacions realitzades per a entrenament i validació.

6.2.4. Predicció

Un cop entrenat l'algoritme, el pas final es aplicar-lo sobre les dades en que no coneixem el resultat i aproximar si els valors obtinguts semblen raonables o no.

Per a fer-ho, modifiquem el codi per a que llegeixi el nou arxiu, prepari el predictor, faci les prediccions sobre els nous valors i els desi en un nou Excel.

7. Anàlisi de resultats

Un cop aconseguit implementar el predictor i utilitzar-lo tant en la fase d'entrenament com de predicció, cal analitzar les dades obtingudes.

7.1.1. Resultats de l'entrenament

Per analitzar l'entrenament, desem els resultats de l'entrenament del predictor contra el subconjunt de validació emprat en la implementació final.

Analitzant-los en detall, observem que s'han predit 1.431 de 10.864 registres com a renovadors mentre que la variable de validació en definia 3.058. Per tant, resulta ser que el predictor simplement opta per actuar "conservador" i donat que la majoria dels registres no son renovadors, aconsegueix el % d'encerts que havíem vist al capítol previ.

Addicionalment, l'algoritme de XGBoost ens permet executar-lo sense modificacions al codi en la seva variant de regressió enlloc de categorització. En aquest format, l'output del programa es un valor entre 0 i 1 (ja que la variable que intentem predir segueix tenint només aquests dos valors possibles) però ens indica la "confiança" de la predicció, es a dir, com de segur està el programa del que està donant com a resultat.

Si extraient aquestes dades, filtrem aquells resultats que tinguin un % de confiança >0,8 per als 1 i menor a 0,2 per als 0, obtenim aquells valors per als que l'algoritme realment està convençut del resultat obtingut.

En fer-ho, resten 5.351 registres (es a dir, el nostre predictor està "segur" en un 50% dels casos) i d'aquests, resulta correcte en un 90% dels casos. El 10% restant resulten ser tots 0s i per tant podríem dir que no ens ha donat en cap cas un client "segur" erroni.

D'aquests resultats en deriven 2 conclusions, per un cantó l'algoritme aconsegueix detectar adequadament gran quantitat de casos i si filem prim agafant només aquells casos "segurs" podem confiar en els resultats. D'altre banda però, hem de tenir en compte que aquest resultat tant positiu es deu a la gran presencia de 0s en la mostra total i que si mirem els 1s, resulta que en la gran majoria dels casos l'algoritme tot i predir-los correctament, ho ha fet sense tenir-ho del tot clar.

7.1.2. Resultats de la predicció

Aplicant el mateix anàlisi que al punt previ, obtenim resultats molt similars (1.382 clients de 10.721 marcats com a renovadors) però com es evident, en aquest cas no podem validar si ha encertat o no.

El que si podem fer però es analitzar la confiança dels resultats emesos i trobem que 5.484 dels 10.721 registres compleixen el mateix llindar que hem usat al punt anterior.

Així doncs, plantegem que en cas d'utilització productiva de l'algoritme com a tal, seria sobretot per a descartar clients que podem "assegurar" que no renovaran ja que com hem vist, en aquests casos l'algoritme ha demostrat una precisió excel·lent.

En el cas dels renovadors, el nostre consell seria d'utilitzar aquells valors en que la confiança de l'algoritme supera el 60% ja que en aquests casos, també resultarà útil.

7.1.3. Lliurable final

Pel que fa el lliurable final, l'únic requeriment serà que l'usuari instal·li python des de la pàgina oficial de https://www.python.org/ on es troba disponible en versió executable per a Windows.

Per a les demés llibreries necessàries, proporcionem a l'annex el codi que desat en un .bat i executat, instal·larà tot el necessari.

Així doncs, si es vol replicar el projecte des de la memòria, només s'han de desar els codis en els formats especificats i executar el .bat seguit del codi en python per al predictor.

Conclusions

Queda satisfet l'objectiu del treball d'aconseguir confeccionar un algoritme predictiu com a cas d'ús pràctic sobre dades reals que permeti a una empresa prendre decisions amb un % de fiabilitat superior que l'anàlisi estadístic de les dades.

Queda demostrat doncs que les eines de Machine learning permeten a les empreses prendre decisions de negoci de manera accelerada, per exemple, arribant a conclusions que treballadors han adquirit en base a anys d'experiència en unes hores de computació.

Aquest tipus d'informació resulta molt potent a l'hora de millorar substancialment en màrqueting, logística i definició d'objectius entre altres però s'ha d'anar en compte ja que una mala praxis pot fer esbiaixar un model molt fàcilment.

Addicionalment, es demostra que per a poder aplicar analítica en el món de les dades, una de les parts més voluminoses en quant a hores i esforç és el tractament, la visualització i l'estudi de les dades per a entendre completament amb què s'està treballant. Si no es realitza aquest tipus de treball previ, resulta impossible extreure informació sense caure en el parany de la diferencia entre correlació i causalitat.

Amb tot això, resulta comprensible que les empreses estiguin avançant cada cop més en aquesta direcció i que els proveïdors d'aquests serveis ofereixin sovint manteniments i ampliacions un cop un client es decideix a entrar al món de l'analítica avançada.

Agraïments

Es necessari agrair a l'empresa SDG Group la possibilitat de realitzar el treball amb ells, facilitant tant el software com la pròpia base de dades en la que s'han fet totes les probes i desenvolupaments així com l'ajuda proporcionada a l'hora d'aprendre tant en l'ús dels aplicatius esmenats al treball com en criteris conceptuals de cara a la correcta explotació de dades.

Bibliografia

Referències bibliogràfiques

[1] Gartner 2019 Magic Quadrant for Data Science and Machine Learning Platforms (as of Nov 2018)

[2] Yan, Ma & Liu, Kang & Guan, Zhibin & Xu, Xinkai & Qian, Xu & Bao, Hong. Background Augmentation Generative Adversarial Networks (BAGANs): Effective Data Generation Based on GAN-Augmented 3D Synthesizing. Symmetry. 10. 734. 10.3390/sym10120734.ASHRAE, American Society of Heating, Refrigerating and Air-Conditioning Engineers, *Fundamentals Volume (S.I. edition.)*. Atlanta: 2001, p. 104-121

[3] Diagrama llicenciat per a la lliure utilització a la Wikipedia sota la llicencia Creative Commons CC0 1.0 Universal Public Domain Dedication https://commons.wikimedia.org/wiki/File:Reinforcement_learning_diagram_fr.svg

[4] Dra. Alexis Cook – Kaggle – Intermediate Machine Learning Course. https://www.kaggle.com/learn/intermediate-machine-learning

Bibliografia complementària

Les referencies esmenades en aquest apartat no s'han referenciat de manera explícita en la memòria però s'han consultat i utilitzat per fonamentar els coneixements reflectits en especial al capítol 4.

[5] Dan Becker – Kaggle – Intro To Machine Learning Course. https://www.kaggle.com/learn/intro-to-machine-learning

[6] Andrew Ng. – Stanford – Coursera – Machine Learning Course. https://www.coursera.org/learn/machine-learning

[7] Saeed Aghabozorgi – IBM – Coursera – Machine Learning With Python Course. https://www.coursera.org/learn/machine-learning-with-python

[8] Victor Roman - Supervised Learning: Basics of Linear Regression https://towardsdatascience.com/supervised-learning-basics-of-linear-regression-1cbab48d0eba

[9] Prashant Gupta - Decision Trees in Machine Learning https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

[10] Sunil Ray - Commonly used Machine Learning Algorithms https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

Per a la definició de Supervised Learning:

[11] Stuart J. Russell, Peter Norvig (2010) Artificial Intelligence: A Modern Approach, Third Edition, Prentice Hall ISBN 9780136042594.

[12] Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar (2012) Foundations of Machine Learning, The MIT Press ISBN 9780262018258.

[13] S. Geman, E. Bienenstock, and R. Doursat (1992). Neural networks and the bias/variance dilemma. Neural Computation 4, 1–58.

[14] G. James (2003) Variance and Bias for General Loss Functions, Machine Learning 51, 115-135. (http://www-bcf.usc.edu/~gareth/research/bv.pdf)

Per a la definició de Unsupervised Learning:

[15] Hinton, Jeffrey; Sejnowski, Terrence (1999). Unsupervised Learning: Foundations of Neural Computation. MIT Press. ISBN 978-0262581684.

[16] Roman, Victor (2019-04-21). "Unsupervised Machine Learning: Clustering Analysis". Medium. Retrieved 2019-10-01.

Per a la definició de Reinforced Learning:

[17] Kaelbling, Leslie P.; Littman, Michael L.; Moore, Andrew W. (1996). "Reinforcement Learning: A Survey". Journal of Artificial Intelligence Research. 4: 237–285. arXiv:cs/9605103. doi:10.1613/jair.301. Archived from the original on 2001-11-20.

[18] Van Otterlo, M.; Wiering, M. (2012). Reinforcement learning and markov decision processes. Reinforcement Learning. Adaptation, Learning, and Optimization. 12. pp. 3–42. doi:10.1007/978-3-642-27645-3_1. ISBN 978-3-642-27644-6.